DOCUMENT RESUME

ED 467 814 TM 034 358

AUTHOR	Schnipke, Deborah L.; Scrams, David J.
TITLE	Representing Response-Time Information in Item Banks. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Princeton, NJ.
REPORT NO	LSAC-R-97-09
PUB DATE	1999-05-00
NOTE	20p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Adaptive Testing; Admission (School); Arithmetic; College Entrance Examinations; *College Students; Computer Assisted Testing; Goodness of Fit; Higher Education; *Item Banks; Law Schools; *Responses; *Test Items; Timed Tests
IDENTIFIERS	*Law School Admission Test

ABSTRACT

The availability of item response times made possible by computerized testing represents an entirely new type of information about test items. This study explores the issue of how to represent response-time information in item banks. Empirical response-time distribution functions can be fit with statistical distribution functions with known properties. For this study, data were obtained from a computerized adaptive test of arithmetic reasoning skills administered as part of a larger test battery (the Law School Admission Test) to 38,357 examinees. Four functions (the normal, lognormal, gamma, and Weibull) were fit to empirical distribution functions from a computer-administered test, and the various functions were evaluated to determine which described the empirical distributions the best and provided the most useful parameters for storing in an item bank. The lognormal distribution was found to best fit both exploratory and confirmatory samples. It provides meaningful and useful parameters that can be stored in an item bank. (Contains 3 tables, 8 figures, and 11 references.) (Author/SLD)

LSAC RESEARCH REPORT SERIES

TM

Representing Response-Time Information in Item Banks

Deborah L. Schnipke and David J. Scrams
Law School Admission Council

Law School Admission Council
Computerized Testing Report 97-09
May 1999

LAW Services

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

As the Law School Admission Council (LSAC) considers computerizing the Law School Admission Test (LSAT), one of the advantages to keep in mind is the availability of item response times, which provide an entirely new type of information about items and test takers. In addition to knowing the accuracy with which test takers answer an item, in computer-administered tests we can investigate the amount of time test takers spend on each item. Researchers have begun to investigate uses of response times for a variety of research topics in the testing field, but to make use of response times in an operational test administration, response-time information about each item needs to be stored in an item bank. Currently there are no guidelines on how to do this. Storing a large number of response-time statistics would be impractical, but simply storing the mean and standard deviation is not sufficient because response times are positively skewed.

The goal of the present study was to develop a method for summarizing response-time information both accurately and concisely. Specifically, we wanted to be able to characterize the entire response-time distribution in terms of a small number of item parameters. Characterizing the entire response-time distribution is necessary so that any desired response-time characteristic can be easily calculated (e.g., median, mean, 95th percentile, dispersion). Doing so with only a few parameters is necessary for easy storage in an item bank.

In this preliminary investigation, we modeled item response times using several statistical distribution functions used by previous researchers to model response times in the testing field. We randomly separated response times to operational items into two samples. We modeled response times in the first sample with the various distribution functions and evaluated the fits. We then fit the models to the second sample using the parameter estimates from the first sample. This allowed us to examine how well the parameter estimates generalize to a new sample. The lognormal distribution provided very good fits for both samples (better than the fits provided by other distribution functions), and the lognormal distribution has only two parameters. Storing these two parameters for every item provides an accurate and concise summary of response-time information.

Abstract

The availability of item response times made possible by computerized testing represents an entirely new type of information about items. In addition to knowing the accuracy with which test takers answer an item, we can now investigate the amount of time test takers spend on each item. The issue of how to represent response-time information in item banks is explored. Empirical response-time distribution functions can be fit with statistical distribution functions with known properties. Four functions (the normal, lognormal, gamma, and Weibull) are fit to empirical distribution functions from a computer-administrated test, and the various functions are evaluated to determine which describe the empirical distributions the best and provide the most useful parameters for storing in an item bank. The lognormal distribution was found to best fit both exploratory and confirmatory samples and it provides meaningful and useful parameters which can be stored in an item bank.

Introduction

One of the advantages of computer-based testing is the availability of response-time information. Researchers have investigated uses of response times for a variety of topics in the testing field. Response times

A version of this paper was presented by David J. Scrams at the annual meeting of the Psychometric Society, 1997, Gatlinburg, TN.

have been used to examine differential speededness across subgroups (O'Neill & Powers, 1993; Schnipke, 1995; Schnipke & Pashley, 1997) and explore relationships between test taker ability and processing speed (Scrams & Schnipke, 1997; Thissen, 1983). Response times have also been used to study pacing during a test (Llabre & Froman, 1987) and identify unusual response patterns (Schnipke & Scrams, 1997). Response times could also be used to establish reasonable time limits (Bhola, Plake & Roos, 1993; Reese, 1993) or to predict finishing times (Roskam, 1997).

For some of these uses, response-time information needs to be stored in an item bank, along with other item characteristics (e.g., difficulty, discrimination, and content codes). The present work is an exploration of how best to characterize items in terms of response times. We wanted a method that accurately described response-time characteristics in a way that allowed for easy incorporation into an item bank. Our general approach is to fit statistical functions with known properties to the empirical response-time functions from an operational test. This approach allows us to characterize the entire response-time distribution using only a small number of statistical parameters.

In this preliminary investigation, we used several distribution functions that have been proposed by previous researchers as response-time models for the testing context: lognormal (Thissen, 1983), gamma (Verhelst, Verstralen, & Jansen, 1997), and Weibull (Roskam, 1997). We also examined the normal distribution function as a standard against which the other functions could be compared. Model fits were examined for both exploratory and confirmatory samples so that generalizability could be evaluated.

Method

Data Characteristics

Data were obtained from a computerized adaptive test of arithmetic reasoning skills administered as part of a larger test battery. The test was fixed length: all 38,357 test takers received 15 multiple-choice items each, from a pool of nearly 200 items[1] Response times were recorded in tenths of a second.

We randomly selected 30 items from the item pool that did not contain graphics and which had at least 1,000 responses. These 30 items were used for all analyses. The number of responses for each of the 30 items ranged from 1,007 to 7,417. A relatively large sample size was important because we randomly divided samples into exploratory and confirmatory samples, and we wanted both samples to be large enough to perform model fits. Table 1 contains summary information about each of the 30 items (e.g., item response theory [IRT] parameter estimates, mean and median response times, and sample size).

1 We do not actually know how many items were in the pool; we only know how many unique items were administered to our sample of test takers.

TABLE 1
Summary information for the 30 items (exploratory and confirmatory samples combined)

	IRT Parameter Estimates				Response Time Characteristics					
Item	a	b	c	Sample Size	Median	Skew	Min	Max	25th Percentile	75th Percentile
1	2.00	0.14	0.34	6,941	46.7	2.112	0.5	471.4	28.4	76.0
2	1.95	0.34	0.25	7,417	55.4	1.935	0.0	433.5	37.1	82.2
3	1.20	-0.63	0.15	1,974	57.4	2.078	-1.3	441.1	36.0	94.1
4	1.39	-0.13	0.22	1,576	46.9	2.442	2.5	443.4	32.1	71.0
5	1.39	-1.10	0.35	2,118	60.3	1.992	3.3	425.5	40.5	90.7
6	1.17	-0.92	0.10	4,437	38.6	1.786	0.4	234.9	27.6	54.8
7	1.61	0.16	0.22	4,401	70.8	1.939	2.0	536.4	46.4	107.7
8	2.61	0.88	0.24	7,551	48.8	2.162	0.0	462.4	32.2	75.2
9	1.77	-0.20	0.29	6,331	53.4	1.998	1.0	361.3	36.9	78.5
10	1.39	0.06	0.17	2,107	56.5	2.212	2.5	411.0	37.7	84.7
11	1.32	-0.03	0.14	3,248	87.6	1.796	6.3	610.6	58.1	133.3
12	1.27	-0.93	0.24	2,050	42.9	2.403	3.4	411.3	28.0	65.8
13	2.61	0.62	0.16	7,265	24.0	2.363	0.0	259.1	16.0	37.0
14	1.65	-0.04	0.14	6,240	40.3	2.051	1.0	313.4	26.9	60.9
15	1.56	-0.10	0.22	6,691	71.9	2.059	0.6	674.7	48.8	112.3
16	2.61	0.95	0.17	8,121	43.0	2.295	0.0	376.2	29.6	65.7
17	1.32	-0.56	0.23	3,582	45.7	2.332	2.3	441.1	29.7	71.3
18	2.10	0.16	0.17	6,567	37.2	2.038	0.8	298.5	25.1	55.0
19	1.29	-0.65	0.14	5,248	56.6	1.831	0.6	531.6	32.4	93.4
20	1.79	-0.22	0.22	6,562	36.7	2.173	1.0	396.4	22.3	61.1
21	2.11	-0.05	0.18	6,621	28.9	2.606	0.7	338.4	18.6	48.2
22	2.36	0.54	0.20	7,022	95.4	1.781	0.0	489.1	72.5	129.3
23	1.15	-1.17	0.17	1,866	65.8	2.017	0.4	487.7	43.1	96.2
24	1.21	-1.10	0.16	3,029	33.9	2.654	0.5	355.0	22.7	53.4
25	1.35	-0.79	0.14	5,681	53.3	2.084	0.4	435.7	36.2	80.0
26	2.57	0.35	0.23	6,777	31.1	3.245	0.0	302.6	24.0	42.7
27	1.53	1.18	0.22	1,007	88.2	1.919	5.7	464.6	65.3	121.7
28	1.65	0.69	0.10	4,774	44.9	2.222	0.9	495.5	26.3	79.0
29	1.83	-0.26	0.25	6,263	26.8	2.620	0.8	275.6	19.7	39.5
30	1.26	-1.09	0.26	1,947	19.5	2.974	3.3	318.0	12.1	35.0

Samples

An exploratory sample was created for each of the 30 items by randomly selecting 500 responses for each item. The remaining responses for each item were considered a confirmatory sample. In all cases the confirmatory sample was larger than the exploratory sample because we only used items which had at least 1,000 responses (in fact, the smallest sample size was 1,007).

Our intent for the exploratory sample was to mimic the pretest situation. Before items are administered operationally on the Law School Admission Test (LSAT), they are "tried out" as pretest items. Pretest items are administered with operational items, but they are not scored; only operational items contribute to a test taker's score. We gather data on pretest items so that we can calculate item difficulty, differential item functioning statistics, and other statistics. If the pretest items are found to be acceptable, they may eventually become operational items. On a computer-administered test, response times can be collected at both the pretest and

operational stage. Response times from the pretest stage can be summarized and stored in an item bank so that this information is available when the items are used operationally.

Distribution Functions

We first fit the exploratory sample for each item with various distribution functions and examined the fits. We then fit the confirmatory sample for each item using the parameter estimates from the exploratory sample for the item and again examined the fits. We used four distribution functions, each of which has two parameters. The first, the normal distribution, was used as a standard for comparison (a yardstick). Researchers and practitioners have used the mean and standard deviation to represent response times, but this practice can lead to confusion because many people readily identify the sample mean and standard deviation with normality. With skewed data, the sample mean is not the best measure of central tendency, and 68% of the data is not within one standard deviation of the mean. Because response-time data tend to be positively skewed, we did not expect the normal distribution to perform very well. We used the normal distribution as a demonstration of this problem. The other three distributions (lognormal, gamma and Weibull) were used because they have been adopted in the context of particular theoretical models of response-time data and because they are unimodal and positively skewed distributions. Thissen (1983) used the lognormal distribution to model response times in his timed-testing model. Verhelst, Verstralen, & Jansen (1997) used the gamma distribution to model response times in speed tests (tests in which speed of performance is essential and where all items are sufficiently easy that they could be answered correctly with high probability by all test takers if time were available). Roskam (1997) used the Weibull distribution to model test completion times (as opposed to individual response times).

The normal density function has location parameter μ and scale parameter σ. The normal density is unimodal and symmetric, unlike response times which are generally positively skewed. The normal probability density function (PDF) for variable t (response time) is given by

$$\mathrm{PDF}_{\mathrm{normal}}(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(t-\mu)^2}{2\sigma^2}\right]. \qquad (1)$$

The parameters of the normal density were estimated by sample statistics: $\hat{\mu} = \bar{t}$ (the sample mean), and $\hat{\sigma} = s_t$ (the sample standard deviation).

The lognormal density has scale parameter μ and shape parameter σ. The lognormal density is unimodal, positively skewed, and always positive, as are response times. The lognormal PDF is given by

$$\mathrm{PDF}_{\mathrm{log\,normal}}(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[\frac{-(\log t-\mu)^2}{2\sigma^2}\right]. \qquad (2)$$

Applying the lognormal density is equivalent to applying the normal density to the logarithm of the raw data. The parameters of the lognormal density were estimated by taking the mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of the natural logarithm of response times.

The gamma density has scale parameter β and shape parameter α. The gamma density is unimodal, positively skewed, and always positive. The gamma PDF is given by

$$PDF_{gamma}(t) = \left(\frac{t}{\beta}\right)^{\alpha-1} \frac{\exp(-t/\beta)}{\beta\,\Gamma(\alpha)}, \qquad (3)$$

where $\Gamma(\alpha)$ is the gamma function with argument α. The parameters of the gamma density were estimated with sample statistics: $\hat{\beta} = s_t^2 / \bar{t}$ and $\hat{\alpha} = (\bar{t}/s_t)^2$.

The Weibull density has scale parameter β and shape parameter α. The Weibull PDF is given by

$$PDF_{Weibull}(t) = (\alpha t^{\alpha-1}/\beta^{\alpha}) \exp\left[-(t/\beta)^{\alpha}\right]. \qquad (4)$$

The parameters for the Weibull density are poorly estimated by sample statistics, so these parameters were estimated by a least-squares fit to the cumulative distribution function[2] (CDF) instead. Because we evaluated the fit of each model based on how well the estimated CDF fit the empirical CDF, the Weibull distribution had an advantage in the exploratory sample over the other three distributions which estimated their parameters with sample statistics.

Results

Sample Empirical Response-Time Distributions

Figure 1 shows a histogram of the response times for item 18[3] (exploratory and confirmatory sample combined), which had a sample size of 6,567. As is typical of response-time densities, the data are unimodal and positively skewed.

2 The cumulative distribution function (CDF) is mathematically equivalent to the probability density function (PDF). The CDF at t is the integral of the PDF from $-\infty$ to t (in the case of the normal distribution) or from 0 to t (in the case of the lognormal, gamma, and Weibull distributions).

3 Item numbers do not reflect item position. Items could potentially be seen in any item position. Items were ranked (best to worst) in terms of the fit of the best-fitting model (lognormal), and the item numbers reflect this ranking.

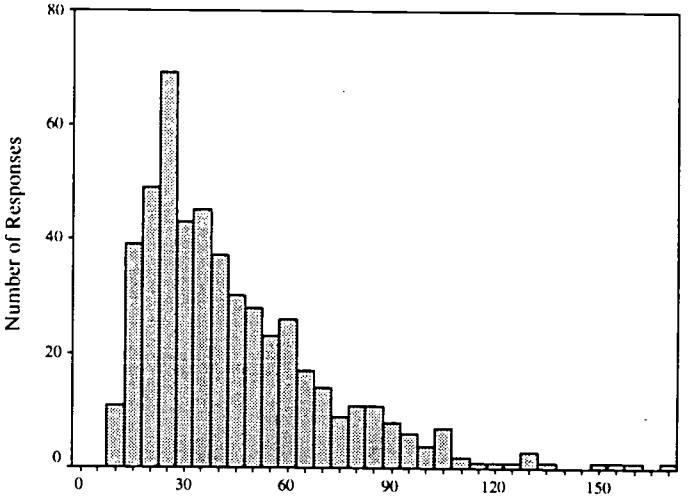

FIGURE 1. *Histogram of response times for item 18. Bars represent 5-second intervals. Item statistics: n = 6,567, mean RT = 44.16 seconds, median RT = 37.20 seconds, SD = 28.11 seconds, skew = 2.04.*

Figure 2 shows a histogram of all response times for item 27 (exploratory and confirmatory sample combined), which had a sample size of 1,007. Again, the data are unimodal and positively skewed. The response times for item 27 are much more spread out than the response times for item 18. The median response time on item 27 was 88.2 seconds, whereas the median on item 18 was 37.2 seconds. One test taker spent more than 7 and a half minutes on item 27, whereas no one took more than 3 minutes to respond to item 18.

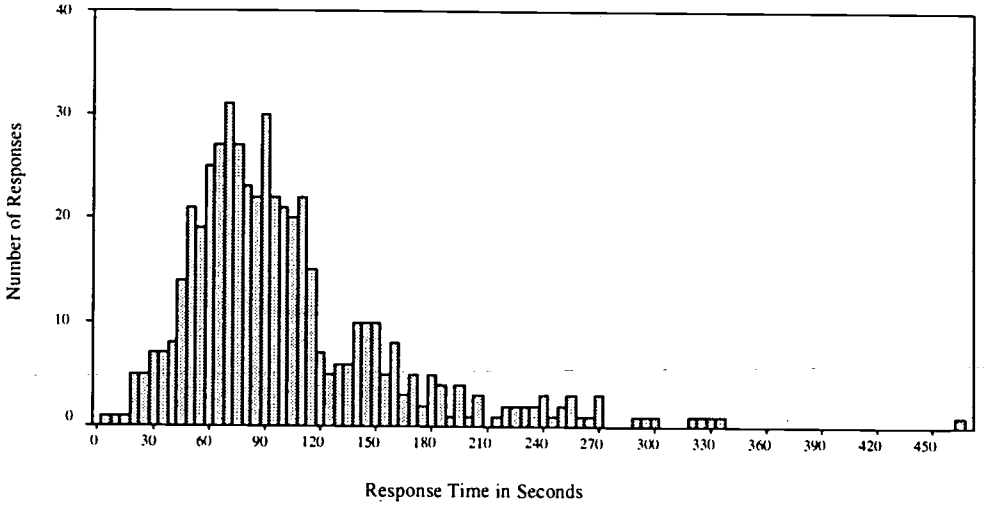

FIGURE 2. *Histogram of response times for item 27. Bars represent 5-second intervals. Item statistics: n = 1,007, mean RT = 103.22 seconds, median RT = 88.20 seconds, SD = 59.90 seconds, skew = 1.92.*

Model Fitting

We first fit the exploratory sample for each item with the four distribution functions. The fits were examined both by visually comparing the expected CDF to the observed CDF for each distribution on each item and by calculating the root mean squared error (*RMSE*; described below) for each distribution on each item. We then fit the confirmatory sample for each item using the parameter estimates from the exploratory sample for the item and again examined the fits in the same ways.

Exploratory Sample

In the first set of analyses, response times for the exploratory sample were fit with the four distribution functions (normal, lognormal, gamma, and Weibull). Summary results will be shown for all 30 items, then detailed results for two items are shown. Recall that the sample size of the exploratory sample was 500 for each item, so comparisons across items are not affected by sample size.

To summarize the fits of the four distributions for each of the 30 items, we calculated *RMSE*. *RMSE* is based on the square root of the mean squared difference between the observed and predicted CDF at every 5^{th} percentile in the observed CDF from the 5^{th} to the 95^{th}. Large values of *RMSE* indicate poor fits.

Table 2 shows the mean *RMSE* for each of the four distributions across items, as well as the minimum (best) and maximum (worst) values of *RMSE*. As shown in Table 2, the lognormal distribution provides the best fit, followed by the gamma, then the Weibull distribution. The normal distribution provides the worst fit overall.

TABLE 2
Summary of RMSE for each distribution in the exploratory sample

	RMSE		
Distribution	Mean	Min	Max
Lognormal	.016	.008	.033
Gamma	.038	.020	.067
Weibull	.051	.030	.076
Normal	.084	.065	.112

To give more detail about how well each distribution fit on each item, the values of *RMSE* for each distribution are shown for each of the 30 items in Figure 3. To make Figure 3 easier to read, items were sorted by the fit of the lognormal distribution because the lognormal provided the overall best fit. These sorted item numbers are used throughout this research report to refer to the items. Items 18 and 27 are used as examples throughout this report.

As shown in Figure 3, the lognormal distribution provided the best fit on every item and the normal distribution provided the worst fit. The gamma and Weibull distributions were in the middle in terms of how well they fit each item, although the gamma provided a better fit than the Weibull on all but two items.

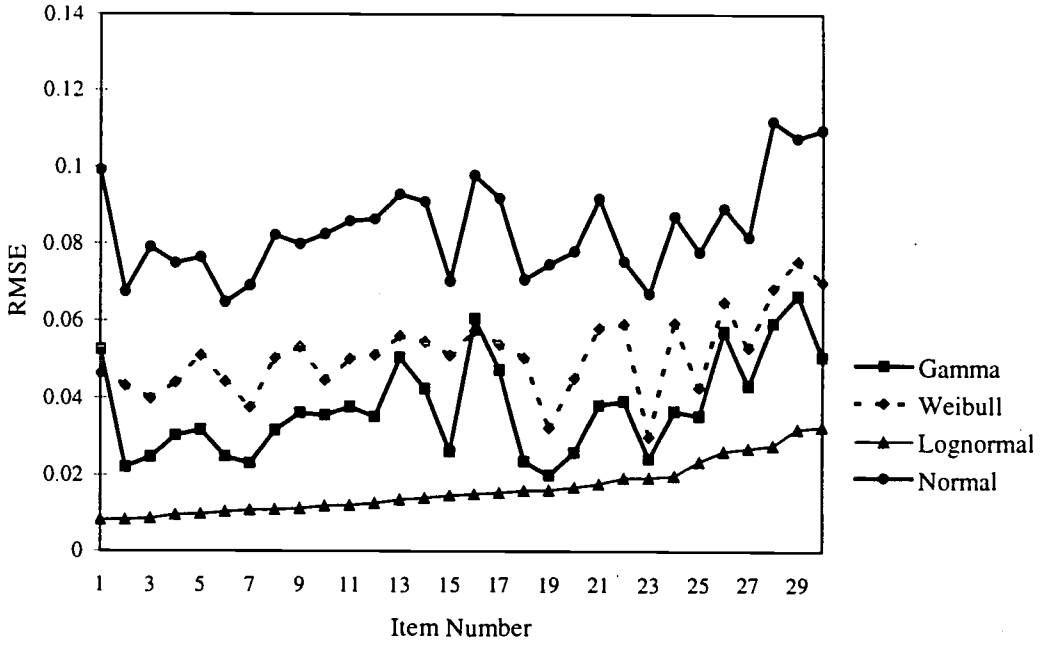

FIGURE 3. *Fit statistics (root mean squared error, RMSE) for each distribution for exploratory sample. The sample size for each item was 500. Item numbers were assigned based on the fit of the lognormal distribution.*

RMSE values indicate the overall fit for each distribution, but they do not provide insight as to why some distributions fit worse than others do. Next we show detailed fits of the four distribution functions to the observed response times for two items: 18 (Figure 4) and 27 (Figure 5). Because items were numbered by the fit of the best model, item 18 is in the "middle of the pack" in terms of fit, and item 27 is one of the worst fit items. These items were selected based on both model fit and sample size. Item 18 had one of the largest sample sizes (6,567), and item 27 had one of the smallest sample sizes (1,007).

Results are shown in the form of double probability plots: the observed cumulative probability is plotted along the abscissa, and the predicted cumulative probability (based on the model) is plotted along the ordinate. A point is plotted for each unique observed response time. The 45° diagonal represents a perfect fit. If the data are fit well by a particular distribution function, the points will cluster around the diagonal.

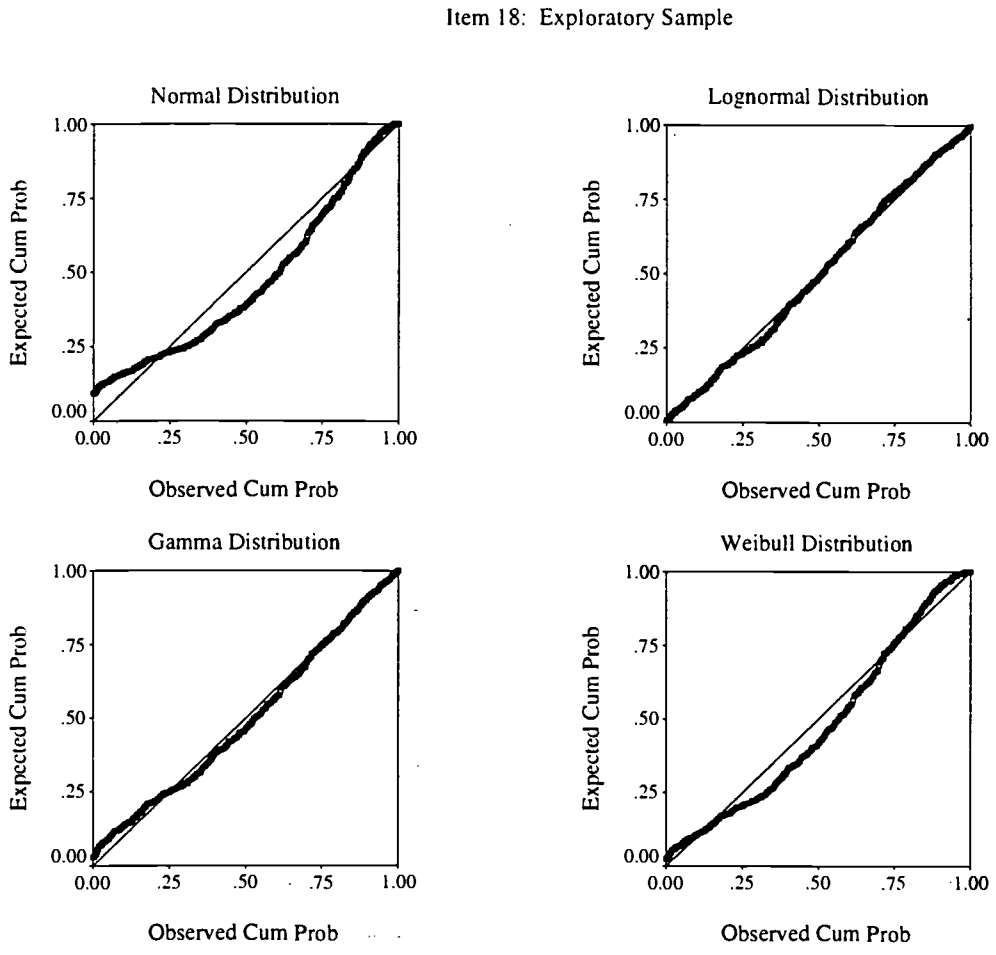

FIGURE 4. *Double probability plots for item 18: exploratory sample (n = 500).*

As shown in the top left panel of Figure 4, the normal distribution provided the worst fit to the response times for item 18. The normal distribution includes negative values (which do not exist in actual response-time data). To account for the skew, the normal distribution has to include many negative response times, so the number of fast response times is overpredicted. Additionally, the skew pulls the mean (44.16 seconds; expected cum prob = .50 in Figure 4) away from the median (37.2 seconds; observed cum prob = .50 in Figure 4). The discrepancy between the median and mean is reflected in the plot by the degree of misfit in the middle of the distribution; the central mass of the observed response-time distribution is located more toward the fast end of the distribution than the normal distribution predicts.

The lognormal distribution fits the observed response times for item 18 quite well, as shown in the top right panel of Figure 4. There are only minor deviations from the diagonal (perfect fit). The gamma distribution (bottom left panel of Figure 4) does not fit quite as well as the lognormal, but the fit is not bad. The gamma distribution slightly overestimates the number of fast response times and slightly misplaces the central mass of the observed response-time distribution.

The Weibull distribution is unable to account for the skew for item 18. As shown in the bottom right panel of Figure 4, the Weibull distribution does not accurately predict the location of the central mass of the observed response-time distribution. The Weibull distribution fits better than the normal distribution, however.

Figure 5 shows the double probability plots for each of the four distribution functions for item 27. The pattern of fits for each distribution on item 27 is similar to the pattern on item 18. As shown in the top left panel of Figure 5, the normal distribution provided the worst fit to the response times for item 27. As in item 18, the amount of skew was a serious problem for the normal distribution: to account for the skew, the normal distribution has to include many negative response times, so the number of fast response times is overpredicted. The skew pulls the mean (103.22 seconds) away from the median (88.2 seconds); thus the normal distribution is not able to predict the location of the central mass of the observed response-time distribution. As in item 18, the central mass of the observed response-time distribution in item 27 is located more toward the fast end of the distribution than the normal distribution predicts.

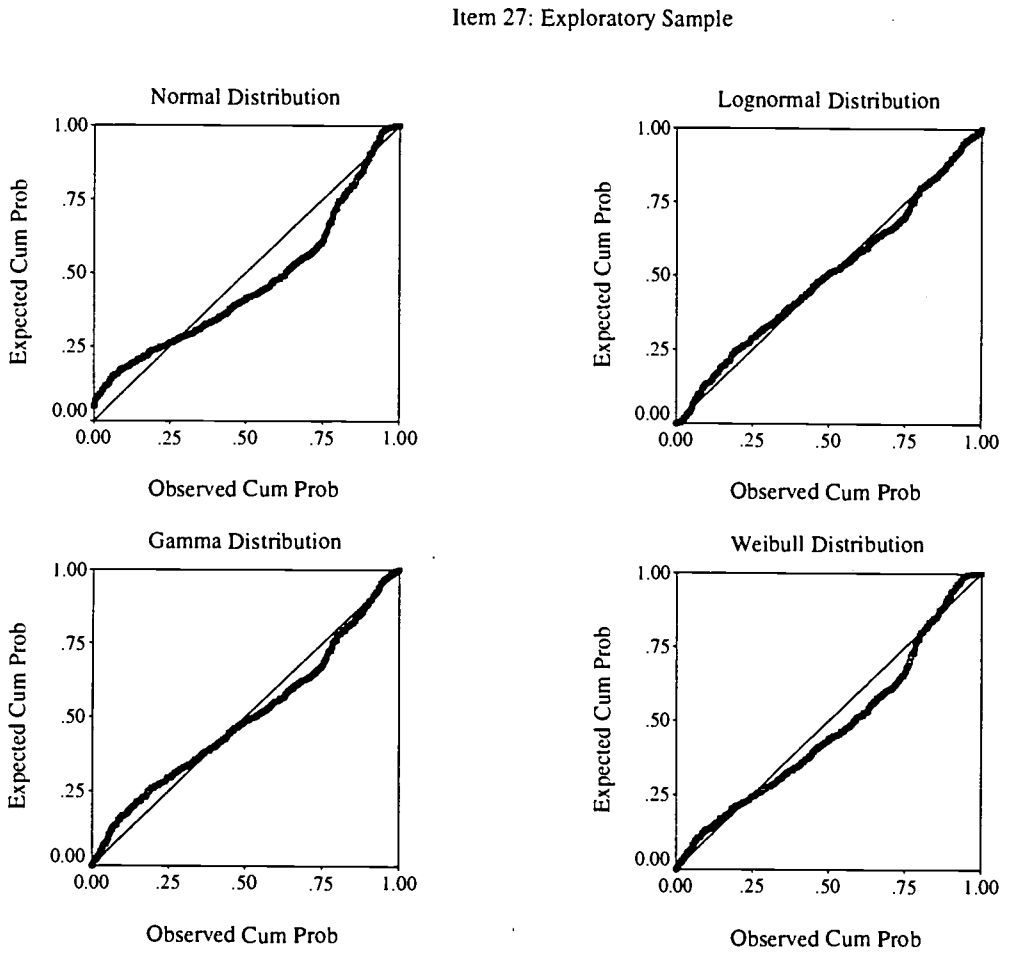

FIGURE 5. *Double probability plots for item 27: exploratory sample (n = 500).*

The lognormal distribution fits the observed response times for item 27 fairly well, as shown in the top right panel of Figure 5. There are only minor deviations. Because the items were ordered by the fit of the lognormal distribution, we know that this was one of the worst fitting items (number 27 of 30 items).

The gamma and Weibull distributions do not fit as well as the lognormal distribution for item 27, but the fit is better than that of the normal distribution, as shown in the bottom left and right panels of Figure 5, respectively. The gamma distribution overestimates the number of fast responses and misplaces the central

mass of the observed response-time distribution. The Weibull distribution also misplaces the central mass of the observed response-time distribution.

Confirmatory Sample

In the second set of analyses, response times for the confirmatory sample were fit with the four distribution functions using the parameter estimates obtained from the exploratory sample. Whereas the sample size of the exploratory sample was fixed at 500 by design, the sample size for items in the confirmatory sample varied from 507 to 6,917.

To summarize the fits for the four distributions for the confirmatory samples, we calculated *RMSE*, as we had done for the exploratory samples. Table 3 shows the mean *RMSE* for each of the four distributions, as well as the minimum (best) and maximum (worst) values of *RMSE*. As in the exploratory sample, the lognormal distribution provides the best fit overall, followed by the gamma distribution, then the Weibull distribution, as shown in Table 3. The normal distribution provides the worst fit overall.

TABLE 3
Summary of RMSE for each distribution in the exploratory sample

	RMSE		
Distribution	Mean	Min	Max
Lognormal	.020	.002	.039
Gamma	.039	.019	.072
Weibull	.049	.026	.075
Normal	.081	.055	.117

Figure 6 shows the values of *RMSE* for the four distributions for each of the 30 items individually for the confirmatory samples. The lognormal provides the best fit on most items for the confirmatory sample, although the gamma and/or Weibull distribution provides a better fit on a few items. The normal distribution provides the worst fit on all items, as it had in the exploratory sample.

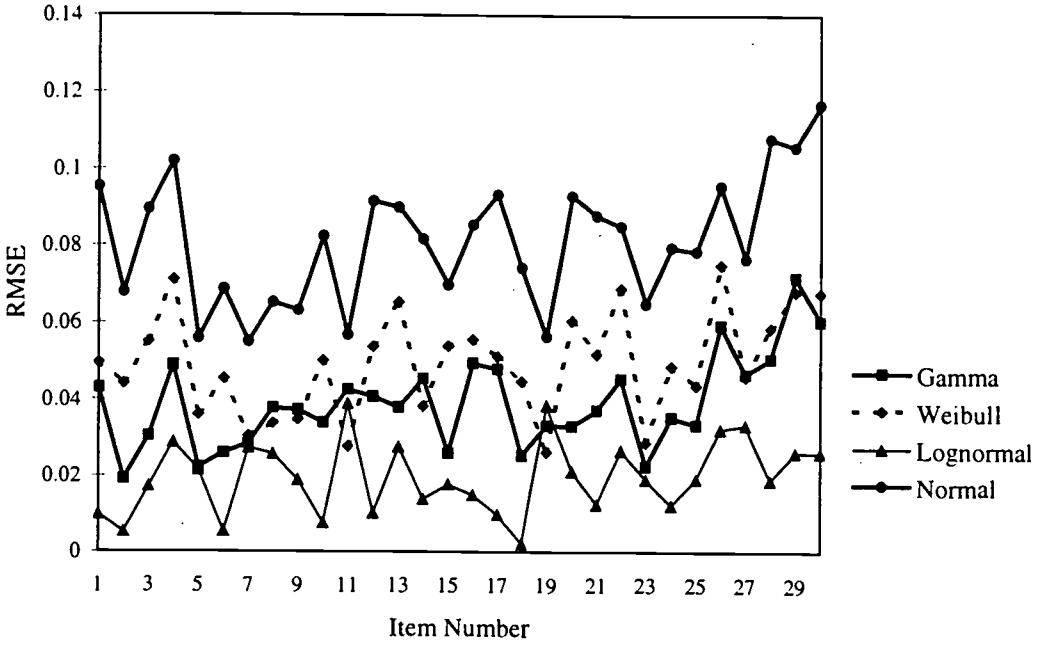

FIGURE 6. *Fit statistics (root mean squared error, RMSE) for each distribution for confirmatory sample, using the parameter estimates from the exploratory sample. Sample sizes range from 507 to 6,917.*

Detailed fits, in the form of double probability plots, are shown for the four distribution functions in Figure 7 for item 18 and Figure 8 for item 27 for the confirmatory samples. In these plots, the parameter estimates from the exploratory sample were used to fit the response times in the confirmatory samples.

In Figure 7, the functions look very smooth because the sample size is large (n = 6,067). As shown in the top left panel of Figure 7, the normal distribution provides the worst fit to the response times for item 18. As in the exploratory sample, to account for the skew, the normal distribution has to include many negative response times, so the number of fast response times is overpredicted. The skew pulls the mean away from the median; thus the normal distribution misplaces the central mass of the observed response-time distribution.

The lognormal distribution fits the observed response times for item 18 very well, as shown in the top right panel of Figure 7. The gamma and Weibull distributions do not fit as well as the lognormal distribution for item 18, but the fit is better than that of the normal distribution, as shown in the bottom left and right panels of Figure 7, respectively.

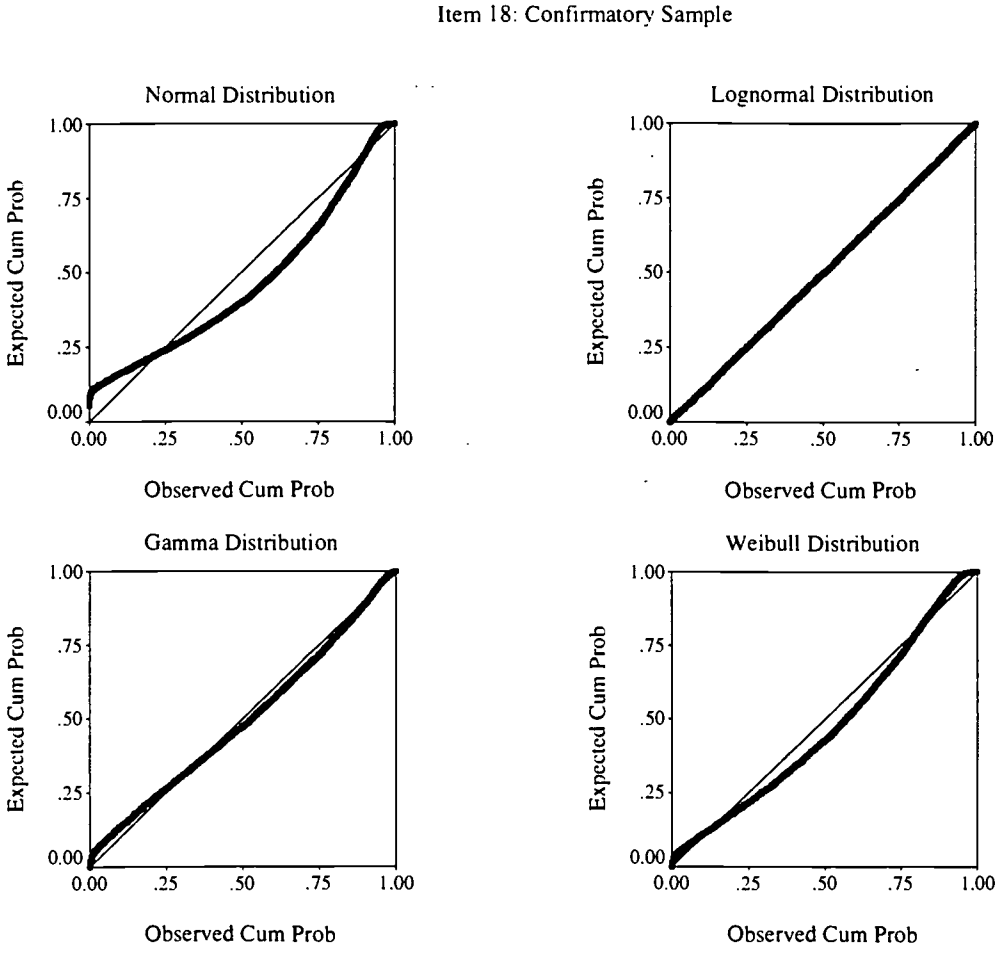

FIGURE 7. *Double probability plots for item 18: confirmatory sample (n = 6,067).*

Figure 8 shows the double probability plots for each of the four distribution functions for item 27 in the confirmatory sample using the parameter estimates from the exploratory sample. The functions for item 27 look less smooth than those for item 18 because the sample size is smaller (n = 507 for item 27). As shown in Figure 8, the normal and Weibull distributions misplace the central mass of the observed response-time distribution. The lognormal distribution fits the observed response times for item 27 fairly well, as shown in the top right panel of Figure 8. The lognormal distribution overestimated the number of fast response times, but fit well elsewhere. The gamma and Weibull distributions do not fit as well as the lognormal distribution for item 27, but the fit is better than that of the normal distribution, as shown in the bottom left and right panels of Figure 8, respectively.

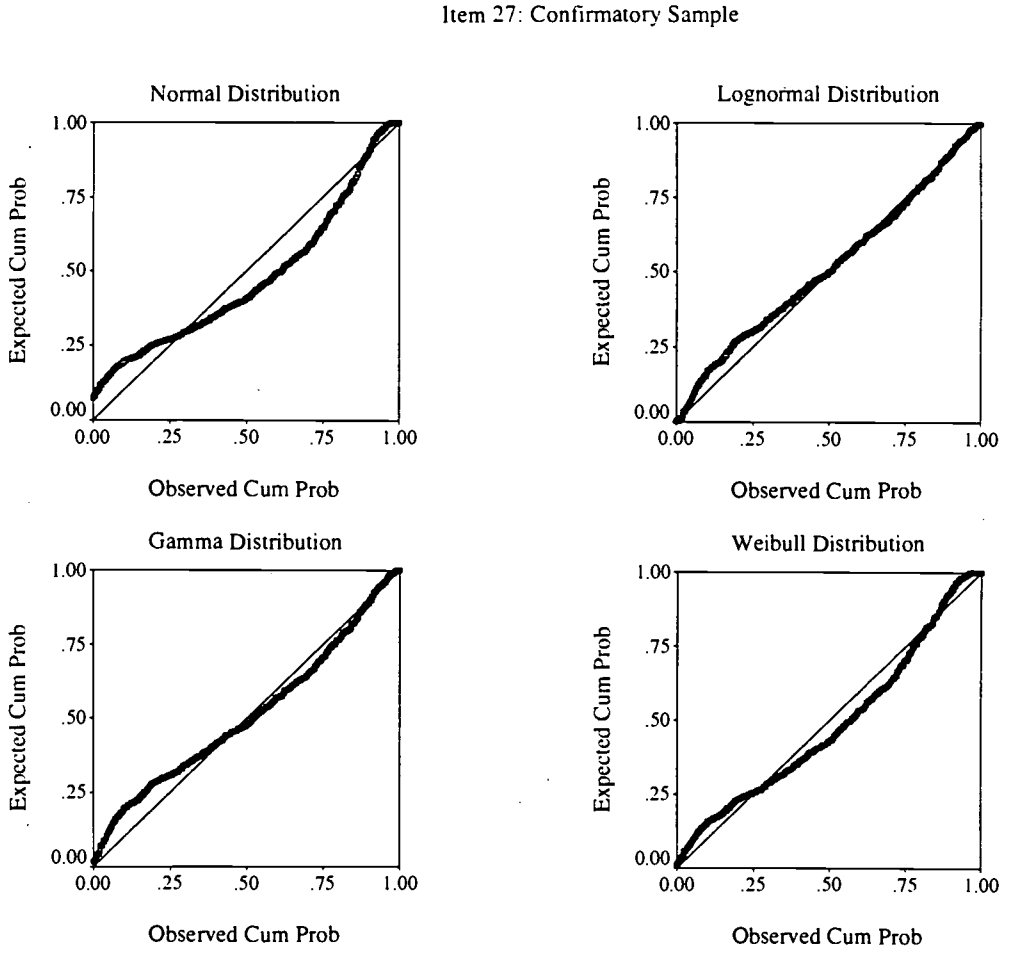

FIGURE 8. *Double probability plots for item 27: confirmatory sample (n = 507; smallest confirmatory sample).*

Discussion

To make use of response times in an operational test administration, response-time information about each item needs to be stored in an item bank. Simply storing the mean and standard deviation is not sufficient because response times are positively skewed. The goal of the present study was to find a way to summarize item response-time distributions accurately and concisely.

We took the approach of fitting response time data with statistical distribution functions as a way of summarizing the entire distribution of response times with a small number of parameters. We used four distribution functions: the normal, lognormal, gamma, and Weibull. Using the normal distribution did not work very well, as expected, because the data were too skewed. The lognormal, gamma, and Weibull distributions are all unimodal and positively skewed and were expected to perform better than the normal distribution. These three distributions have also been used by other researchers to model response times in testing, providing an additional reason for us to try these distributions. We found that the Weibull distribution could not deal with the amount of skew for most items, although it fit better than the normal distribution. The gamma distribution fit a little better on average than the Weibull distribution. The lognormal distribution fit the best of all and, in fact, provided a very good fit for most items.

Besides providing the best fit of response time data, the lognormal also has relatively intuitive parameters. The parameters of the lognormal distribution are the mean and standard deviation of the natural logarithm of original values, so the parameters are easily estimated from sample statistics.

By storing the mean and standard deviation of the logarithm of the response times (the parameters of the lognormal distribution), we can recover the entire distribution of response times for each item quite well. We found that the parameter estimates from the first sample (the exploratory sample) performed well in a cross-validation sample (the confirmatory sample). This suggests that it may be possible to collect data from a pretest sample (equivalent to our exploratory sample) and use the parameter estimates to predict response-time characteristics when the items are administered operationally (equivalent to our confirmatory sample).

In the present work, we used samples of 500 in the exploratory sample. Additional research will take samples of different sizes to see how small the exploratory sample can be and still reasonably predict the remaining responses. This is related to seeing how large a pretest sample needs to be to gather enough response-time data for storing in the item bank. We would also like to take multiple exploratory samples, rather than just one as in the present study. This will allow us to investigate the standard errors associated with these procedures.

Finally, additional research is needed on how ability impacts response-time distributions and how ability distributions of the test takers differ between pretest and operational samples. If ability impacts response-time distributions, pretest response times cannot be used directly to predict operational response times if the ability distribution of the test takers who take pretest items differs from the ability distribution of the test takers who take operational items. However, if the relationship between ability and response time can be determined for an item type, pretest data may be altered for use operationally. (That is, we may be able to correct for ability effects and differences.)

References

Bhola, D. S., Plake, B. S., & Roos, L. L. (1993, October). *Setting an optimum time limit for a computer-administered test*. Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago.

Llabre, M, M., & Froman, T. W. (1987). Allocation of time to test items: A study of ethnic differences. *Journal of Experimental Education, 55*, 137-140.

O'Neill, K., & Powers, D. E. (1993, April). *The performance of examinee subgroups on a computer-administered test of basic academic skills.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Reese, C. M. (1993, April). *Establishing time limits for the GRE computer adaptive tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187-208). New York: Springer.

Schnipke, D. L. (1995). Assessing speededness in computer-based tests using item response times (Doctoral dissertation, Johns Hopkins University, 1995). *Dissertation Abstracts International, 57*, B759.

Schnipke, D. L., & Pashley, P. J. (1997, March). *Assessing subgroup differences in item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213-232.

Scrams, D. J., & Schnipke, D. L. (1997, March). *Making use of response times in standardized tests: Are accuracy and speed measuring the same thing?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169-185). New York: Springer.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").